Future Generation Computer Systems 36 (2014) 156–169

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

An optimal and stable feature selection approach for traffic classification based on multi-criterion fusion

Adil Fahad [a,c,*], Zahir Tari [a], Ibrahim Khalil [a], Abdulmohsen Almalawi [a,d], Albert Y. Zomaya [b]

[a] School of Computer Science and Information Technology, RMIT University, Melbourne, Victoria, Australia
[b] School of Information Technologies, The University of Sydney, Sydney, NSW, Australia
[c] Department of Computer Science, Al-Baha University, Al-Baha City, Saudi Arabia
[d] Faculty of Computing and IT King Abdulaziz University, Jeddah, Saudi Arabia

HIGHLIGHTS

- Propose two new metrics to evaluate the existing FS for network traffic classification.
- Propose a robust multi-criterion fusion-based to preserve the optimal and stable features.
- Propose an adaptive threshold based maximum *entropy* to extract the stable features.
- Propose a wrapper method based on *random forest* to obtain the final optimal subset.
- Improve the performance of traffic classification across different period and networks.

ARTICLE INFO

Article history:
Received 15 March 2013
Received in revised form
25 August 2013
Accepted 5 September 2013
Available online 24 September 2013

Keywords:
Feature selection
Metrics
Internet traffic
Network security

ABSTRACT

There is significant interest in the network management community about the need to identify the most optimal and stable features for network traffic data. In practice, feature selection techniques are used as a pre-processing step to eliminate meaningless features, and also as a tool to reveal the set of optimal features. Unfortunately, such techniques are often sensitive to a small variation in the traffic data. Thus, obtaining a stable feature set is crucial in enhancing the confidence of network operators. This paper proposes an robust approach, called the Global Optimization Approach (GOA), to identify both optimal and stable features, relying on multi-criterion fusion-based feature selection technique and an information-theoretic method. The proposed GOA first combines multiple well-known FS techniques to yield a possible optimal feature subsets across different traffic datasets; then the proposed adaptive threshold, which is based on entropy to extract the stable features. A new *goodness* measure is proposed within a Random Forest framework to estimate the final optimum feature subset. Experimental studies on network traffic data in spatial and temporal domains show that the proposed GOA approach outperforms the commonly used feature selection techniques for traffic classification task.

1. Introduction

Network traffic classification has attracted a lot of interest in various areas [1,2], including Supervisory Control and Data Acquisition (SCADA) [3] security monitoring, Internet user accounting, Quality of Service(QoS), and user behaviour. Classification-based techniques [1,4,5] rely on a set of optimal features to develop accurate and realistic traffic models. The identification of such features for classification purposes is a challenging task because of the following: (i) it requires expert knowledge of the domain to understand which features are important; (ii) the datasets may contain redundant and irrelevant features (that greatly reduce the accuracy of the classification process); and (iii) the efficiency of the classifiers is affected when analysing a large number of features.

Previous classification approaches [6,7] used basic information from IP headers and payload (such as the packet content) for classification. These did not work well, as IP headers contained a few features (e.g. IP addresses, port numbers, protocols) that failed to accurately distinguish between applications. Payload-based techniques [8,9] rely on deep inspection of packet content, which obviously result in significant amount of processing and memory requirements. Recent classification techniques [1,10,11]

* Corresponding author at: School of Computer Science and Information Technology, RMIT University, Melbourne, Victoria, Australia.
E-mail addresses: alharthi.adil@gmail.com, aalharthi.ahmed@gmail.com (A. Fahad).

0167-739X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.future.2013.09.015

deal with such limitations by: (i) avoiding deep packet inspection and thereby creating additional new features from Transport Layer Statistics (TLS) (e.g. packet length and packet arrival time); and (ii) applying specific machine learning techniques [12,13,5] to learn from the data. However, classifying Internet traffic data using TLS is a daunting task in network traffic analysis, as the presence of relevant/redundant features in the traffic data degrades the predictive accuracy of the classifier, maximises the training and testing processing time of the classifier, and finally increases its storage requirements [14]. With the aim of choosing a subset of good features, feature selection techniques (FS) can play an effective role in reducing the dimensionality (of the feature set) and removing irrelevant and redundant features [15].

1.1. Problem statement

Despite a vast number of FS techniques proposed in the literature [16,1,5], ascertaining the benefits of one technique over another still remains a challenge, as each technique is designed with different evaluation criteria (see Section 2.2 for details). Thus, one of the key challenges (in selecting appropriate FS techniques) is the lack of *metrics* to properly compare the existing FS techniques for network traffic classification.

Another key challenge (for traffic classification) is that many FS techniques [16,1,12,8] have been developed with a focus on improving accuracy and performance by discarding the relevant and/or redundant features. However, these studies neglected the insensitivity of the output of FS techniques to variations in the training dataset. This characteristic is referred to here as the *stability* of features. For example, a given FS technique may select largely different subsets of features under small variations of the traffic training data. However, most of these selected features are as good as each other in terms of achieving high classification accuracy and better efficiency. Such an instability issue dampens the confidence of network operators in relying on any of the various subsets of selected features (that can change radically on datasets taken over a period of time). Nevertheless, unstable features in traffic applications are problematic because there is no prior knowledge about the data and therefore, in most cases, these features are subsequently analysed further, requiring much time and effort.

Apart from identifying the stable and optimal features for traffic classification, transport layer statistics (TLS) involve several continuous-valued features. Examples of such features include the number of packets, number of bytes, and duration for each connection. As a consequence, these features can have a negative impact on some machine learning algorithms, in terms of both accuracy and/or training time [17].

1.2. Contributions

This paper deals with the issues described above, and it proposes a new FS technique as well as a discretisation algorithm to enhance the capabilities of the network classification task.

The significant contributions of this paper are:

- two new metrics, namely *optimality* (to measure the quality of the generated feature set by each FS technique) and *stability* (to measure the sensitivity of a FS technique under variations to the training traffic data). Section 2 provides the details. Such metrics are important (i) to jointly study various FS techniques, (ii) to gain a deeper understanding of some successful techniques, and (iii) to derive novel approaches with better performance. A preliminary comparative study of FS techniques is given in Section 2.3, and it suggests that each selected FS technique has its own advantages and *no* single technique performs well across all the metrics.
- a general framework that not only provides the optimal features, but it also automatically discovers the stable features for network traffic. For this purpose, the proposed GOA technique proceeds in *three* phases. The first phase combines the multiple FS techniques to yield the optimal feature subsets across different traffic datasets. In the second phase, instead of relying on a fixed threshold, GOA adapts the concept of *maximum entropy*[1] [18] that culls stable features based on feature distribution. Intuitively, features with a distinct distribution are considered to be stable and are therefore extracted. This process automatically adapts to the feature distribution (i) to yield feature subsets with a distinct distribution (with highest distribution) and (ii) to help narrow the scope for a deeper investigation into specific features set (Section 3.2 for details). In the third phase, the extracted features (obtained from the first and second phases) are passed to a more computationally intense procedure, called *Random Forest* filtering, to determine the most representative features that are strongly related to target classes (e.g. WWW, FTP, Attack). The feature subset with the highest goodness is chosen as the final optimal sets for network classification (Section 3.3).
- optimal discretisations produced by *entropy minimisation heuristics* method [19]. The necessity of using such a method on traffic data can have many reasons. Many machine learning (ML) algorithms primarily handle nominal features [20–22], or may even only deal with discrete features. Even though ML algorithms can deal with continuous features, learning is less efficient and effective [20]. Another advantage derived from discretisation is the reduction and simplification of data which makes the learning faster and produces a more accurate, compact and smaller output. Also, the noise present in the traffic data is reduced. In particular, features discretisation involves partitioning the range of continuous-valued features into a set of mutually exclusive intervals with interval boundaries such that the loss of class/attribute inter-dependence is minimised (Section 5.2).

The proposed work is evaluated using publicly available benchmark traffic datasets collected from a variety of links at the core Internet [23]. We compare the effectiveness and efficiency of the candidate features set against two well-known techniques, namely FCBF–NB [5] and BNN [1]. Extensive experiments show that GOA can identify a smaller set of optimal features compared to the aforementioned FS techniques, with up to about 50%–75% reduction in the feature set size (without degrading the quality of obtained subset). The experimental results also show that, by using the candidate features classification, efficiency is improved by a factor of 50% compared with the average efficiency of FCBF–NB and BNN. Additionally, GOA can identify different traffic applications and anomalous behaviour greater than 96% in the worst circumstances, in both the temporal domain: comparing across different period of time, and the spatial-domain: comparing across different network-locations.

1.3. Organisation of the paper

The rest of this paper is organised as follows. Section 2 briefly reviews some well-known FS techniques and also analyses their performance according to proposed metrics. Section 3 describes the GOA approach, and Section 4 shows the various performance results with the various benchmark datasets. We conclude with some remarks in Section 6.

[1] It is a general technique for estimating probability distributions from the data.